Collages as Dynamic Summaries of Mined Video Content for Intelligent Multimedia Knowledge Management

Tobun D. Ng Michael G. Christel Alexander G. Hauptmann Howard D. Wactlar

Carnegie Mellon University Computer Science Dept.
5000 Forbes Ave., Pittsburgh, PA 15213 USA
{tng, christel, hauptmann, wactlar}@cs.cmu.edu

Abstract

The *video collage* is a novel effective interface for dynamically summarizing and presenting mined multimedia information from video collections. We will discuss how collages are automatically produced, illustrates their use, and evaluates their effectiveness as summaries across news stories. Collages are presentations of text and images extracted from multiple video sources. They provide an interactive visualization for a set of analyzed video documents, summarizing their contents and offering a navigation aid for further exploration. The dynamic creation of collages is based on user context, e.g., an originating query, coupled with automatic processing to refine the candidate imagery. Named entity identification and common phrase extraction provides associative textual description. The dynamic manipulation of collages allows user-directed filtering as well as reveals additional detail and previously undiscovered content. The utility of collages as summaries is examined with respect to other published news summaries.

Introduction

The Informedia Project at Carnegie Mellon University has created a multi-terabyte digital video library consisting of thousands of hours of video, segmented into tens of thousands of documents. Since Informedia's inception in 1994, numerous interfaces have been developed and tested for accessing this library, including work on multimedia abstractions or *surrogates* that represent a video document in an abbreviated manner (Christel 1999) (Wactlar et al. 1999). The interfaces, including video surrogates, build from automatically derived descriptive data, i.e., metadata, such as transcripts and representative thumbnail images derived from speech recognition, image processing, and language processing. This paper introduces the *video collage*: a surrogate representing a summary of multiple video documents through extracted metadata text, images, audio and video as well as a navigation interface for drilling down in various levels of abstraction according to users' information needs. The work presented here focuses on text and images in collage construction and use.

Collages are needed for multimedia knowledge management because for large video libraries, the simple information seeking process of "successively refining a query until it retrieves all and only those documents relevant to the original information need" (Hearst 1999, p. 263) rarely applies. User queries produce hundreds of video documents, and traversing through a linear list of these documents is too time-consuming for users to repeatedly refine their queries. For large video collections, answers are often found across documents, and the "berry-picking" model of information seeking is more likely, which acknowledges that users learn during the search process. This model from Bates (Bates 1989), is expressed as two main points by Hearst (Hearst 1999):

- As a result of learning from the information encountered throughout the search process, the users' information needs continually shift.
- Users' information needs are not satisfied by a single, final set of documents, but rather by a series of selections and bits of information found along the way.

Collages are new interactive tools facilitating efficient, intelligent browsing of video information by users as they follow their shifting information needs. Collages can also be briefing tools presenting many facets of information all at once to communicate a summary across documents to satisfy a particular purpose. These two facets of collages, as interactive summaries and summary snapshots, are the focus of Section 4 and Section 5, respectively. Section 2 discusses the automatic analysis producing data that feeds into collages, and Section 3 outlines how this data is reduced to a manageable set and combined for greater effectiveness.

Video Content Analysis

The traditional video hierarchy for digital video collections has been the decomposition of source video into documents (i.e., stories), and documents into shots, typically through color histogram changes (Wactlar et al. 1999) (Zhang et al. 1994). In additional to retaining structural and logical analysis for segments and shots, Informedia processing adds a variety of multimedia content analysis capabilities. It generates image metadata in the form of identified news

anchorperson shots, face detection, and semantic-bearing features such as indoor and outdoor scenes. It also derived video metadata in the form of camera motion. Regarding text metadata, it adds transcripts via capturing video closed-captioning or generating it through speech recognition, determining when each word was spoken through an automatic alignment process using the Sphinx-III speech recognizer, and filtering the text into a mixed upper and lower case presentation. Further text metadata for the video document comes from detecting and extracting text overlaid on the video image (Wactlar et al. 1999).

Text metadata is used as source material for building word search indices, deriving named entities, and extracting phrases. The word search service returns an inverse document frequency (idf) metric for matching terms, so that following a text query, the contribution of each matching word within a video document to that document's relevance score can be examined. For example, following a query on "government execution", an idf-based metric returns a much higher score for the more unique word "execution", so a match on "execution" within a shot may be more significant to a document's query relevance than two matches to "government" within another shot.

Named entity extraction of people, organization, and location from broadcast news speech transcripts has been done by MITRE via Alembic (Merlino, Morey, and Maybury 1997), and BBN with Nymble (Bikel et al. 1997) (Miller et al. 1999). Similarly, our processing starts with training data where all words are tagged as people, organizations, locations, or something else. We use a statistical language modeling toolkit (Clarkson and Rosenfeld 1997) to build a tri-gram language model from this training data, which alternates between a named-entity tag and a word, i.e. **–none-** *here's* **–person-** *Wolf* **–person-** *Blitzer* **–none-** *for* **–organization-** *CNN* **–none-** *in* **–location-** *Kabul*. To label named entities in new text, we first build a lattice from the text where each text word can be preceded by any of the named-entity tags. A Viterbi algorithm then finds the best path through the named-entity options and the text words, just like speech recognition hypothesis decoding.

Natural language processing (NLP) tools such as Brill's part-of-speech tagger (Brill 1992) have been used for extracting noun phrases from text collection. However, such NLP tools work well with grammatically and syntactically correct English sentences, which are also well formed with proper capitalization and punctuation. Unfortunately, text metadata in video are errorful and ill formed: errors in speech recognition, OCR, spelling, and punctuation as well as non-capitalized and incomplete text stream. Thus, we use a heuristic-based, automatic phrase formation method to extract phrases. Our phrase extraction process first delineates text using common verbs and words, as well as punctuations as delimiters, and then forms phrases with up to 5 adjacent words.

While the Informedia corpus includes broadcast news, documentaries, classroom lectures, and other video genres, this paper focuses on interfaces for broadcast news. Video preview features that work well for one genre may not be suitable for a different type of video (Li 2000), and while the discussion here for news may apply equally well for visual genres like travel and sports videos where the audio also contains an information-rich, well-synchronized narrative, other genres like classroom lecture and conference presentations may need to emphasize speaker, dialogue text, and other unique attributes.

Preparing News Video Collages

A CNN 2001 news corpus is the focus for the remainder of this paper. The corpus used contains 630 CNN broadcasts with a total duration of 459 hours, segmented into 20,744 video documents. A first step in reducing the corpus complexity is to flag documents that are irrelevant to most purposes, which after pilot studies were found to be commercials and weather reports. We developed a commercial detector leveraging from past work (Wactlar et al. 1999) (Zhang et al. 1994), based in part on black frames, shot frequency, speech recognition output and document length. We developed a weather report detector by using weather text stories and images of weather maps to train a Support Vector Machine that classifies documents as weather reports. Removing the detected commercials and weather reports reduced the corpus from 20,744 to 9625 video documents (333 hours), which is our working set.

In this set, there are 162,485 shots, an average of 17 shots per video document. The average shot duration is 7.3 seconds, the average document duration is 2:04.5, and the average transcript size is 1605 bytes. 61,182 mentions of people, places, and organizations were found in the documents through automatic named entity extraction. These numbers illustrate the need to reduce text and image complexity when summarizing hundreds of documents drawn from the collection: aggregating all the data for the documents would take too long for a person to sift through for relevant points.

Reducing the Image Working Set

Even within a single year of news, queries can return an overwhelming amount of data: the most relevant 1000 documents returned by the query "terrorism" contain 17545 shots, while the top 1000 documents for the query "bomb threat" return 18,804 shots. For a collage to draw from these huge sets of shots, we make use of visual significance and query significance. We currently employ only one factor for visual significance in the news: nonanchor shots

are more visually interesting and should get preference over anchorperson shots showing a person behind a desk in the CNN studio.

An anchorperson shot detector, whose first use in filtering digital video news dates back to 1994 (Zhang et al. 1994), was developed based on color histogram areas, using a Support Vector Machine for training and classification. The training set consisted of 4500 instances, 660 of which were identified anchors. From our working set, 10,222 anchorperson shots were automatically detected.

Query significance has been shown to produce thumbnail surrogates that effectively represent a single video document and result in faster information retrieval times (Wactlar et al. 1999). Following a query, the matching terms for the query are identified, and synchronization metadata used to find the highest scoring shot for the document based on idf metrics. The highest scoring shot's thumbnail image representation is used to represent the document. By extending this concept across multiple documents, each document could be represented by its highest scoring shot, or by its highest scoring nonanchor shot if one exists. So, for a query set of 1000 resulting documents, there are 1000 representative images, with nonanchor shots favored over anchor shots.

Reducing the Text Working Set

The transcript text is too large to use completely: a result set of 1000 documents would have over a million bytes of text. Even for a single video document, representing that document with extracted phrases rather than the full transcript text has benefit, when accompanied by shot images (Christel and Warmack 2001). Past investigations into titles, storyboards, and skims found that phrases rather than words were good building blocks for surrogates (Christel and Warmack 2001) (Wactlar et al. 1999). Therefore, transcripts are decomposed into sets of phrases for subsequent use in the video collage.

Given a set of 1000 documents, the phrases that are most common in the transcripts of these 1000 documents can then be returned. In addition, the named entities that are common across documents in a result set can be listed, grouped into categories of person, location, and organization. The number of phrases and named entities, labeled henceforth as "terms", to draw into a collage is controlled through user-settable parameters:

- Maximum number of terms to show
- Minimum number of documents that the term appears in, as an absolute value, e.g., term must appear in 3 or more documents
- Minimum percentage of documents in a set that the term must appear in, e.g., term must appear in 1% of documents (so for a collage of 500 documents, term must appear in 5 or more documents)

The settings used in collages presented in later sections are: maximum of 10 terms (sometimes cropped further in figures), occurring in at least 2 documents, ordered by the number of documents containing that term (percentage filter not used).

Combining Text and Imagery

Based on prior studies that have shown that the presentation of captions with pictures can significantly improve both recall and comprehension, compared to either pictures or captions alone (Large 1995), the combination of text and visuals in a collage should be better than using solely text or images. Indeed, Ding et al. found that video surrogates including both text and imagery are more effective than either modality alone (Ding et al. 1999). This work was confirmed in a recent study (Christel and Warmack 2001), which specifically examined the questions of text layouts and lengths in storyboards. Text phrases coupled with imagery produced a useful surrogate for efficient information assessment and navigation within a single video document. Collages were hence used to employ both images and text. The next section illustrates collage surrogates of multiple video documents.

Interacting with Collages

The elements in the Informedia digital video library interface serve both to communicate information efficiently as well as to facilitate browsing and navigation. For example, the map interface is used to show distribution of geographic regions within a particular video document as well as across documents, and also for issuing a query to a chosen set of countries or selected geographic area (Christel 1999). In similar fashion, the video collage serves to communicate a summary for a set of data, while allowing the user to "drill down" into subsets of video to expose greater detail for those areas of interest. Collages were designed to support Shneiderman's Visual Information Seeking Mantra (Shneiderman 1996): "Overview first, zoom and filter, then details-on-demand." Collages as overviews allow the rich information space covered by a set of news video documents to be better appreciated and understood. The user can then "zoom in" to focus points of interest, such as documents containing particular query words, certain geographic areas via map interaction, or specific time occurrence via timelines (Christel 1999), getting immediate feedback within a fraction of a second. Prior work in interacting with query terms, maps, and timelines has been extended with the introduction of thumbnail images and text phrases into the

presentation. This section illustrates how these additional attributes improve the utility of the collage when used in an interactive fashion.

Shahraray notes that "well-designed human-machine interfaces that combine the intelligence of humans with the speed and power of computers will play a major role in creating a practical compromise between fully manual and completely automatic multimedia information retrieval systems" (Chang 1999). The power of the collage interface derives from its providing a view into a set of video documents, where the user can easily modify the view to emphasize various features of interest.

Map Collage Example

Consider a user interested in finding geographic distribution on reports of political asylum and refugees in 2001 news. Following a query on these words, 126 video documents are returned. At an average of 2 minutes per document, the user would need over 4 hours to examine all of the material, possibly without retaining an overview of how these stories relate to countries in Africa and the Middle East. If this became the subject of inquiry, the user could open up a map collage and use the standard map interface tools of pan and zoom to select to see this region of interest, as shown in Figure 1.

Figure 1. Map collage, with common phrases and frequent locations for documents pertaining to Africa.

The collage includes images tiled over the map: images taken from the matching shots of the most relevant documents discussing a particular country. Also shown are lists of the most common phrases and most frequently mentioned locations, via named entity extraction operating on the transcript, for the documents plotted against the map. When the user focuses on all of Africa and the Middle East, stories about the Israeli-Palestinian conflict and Afghanistan-Osama bin Laden are evident. When the user focuses in on West Africa, as shown in Figure 2, the number of images for each country can increase as its plot area increases, thereby showing more visual detail for that country. Each image can be made larger, and the descriptive text changes, in this case indicating that refugee stories in this area of the world for 2001 deal with Sierra Leone and Mali, with a theme on an international movement of people from Africa to Europe, sometimes illegally.

A tiling metric is used to overlay the images on the map collage. A grid is overlaid on the map, and images are displayed in each grid tile intersecting a scoring region until a threshold set by the user is reached, with one image per tile, the highest-scoring regions reserving tiles first. Advantages of this approach include images that don't obscure each other, and images are located by the countries they describe. The obvious disadvantage to this approach is that countries with large areas have greater potential to show more images. The overlapping image view used in timeline collages will be investigated with maps as well.

Figure 2. Map collage after zooming into a subset of Africa from the collage shown in Figure 1.

Manipulating Collages Dynamically

Under user control, the collage becomes an interactive summary, highlighting details for the user's selected region of interest. Figure 2 shows a geographic area receiving greater focus. Figure 3 illustrates the use of dynamic query sliders (Ahlberg and Shneiderman 1994) to select a date

range following a map query on Israel and its immediate neighboring countries. 409 video documents are returned from the map query, and the timeline collage shows these documents plotted by their broadcast date along the x-axis with the y-axis indicating relevance to the query. The user can adjust the date slider to reduce the time interval to two-month blocks, and then drag that two-month window from one endpoint of the slider to the other to show story distributions, counts, frequent named entities, common phrases, and representative images for the active intervals. Sliding across the month pairs January-February, February-March, March-April, April-May, May-June, etc., shows the story count for these intervals reduced from 409 to 55, 46, 28, 49, 63, etc. Figure 3 shows 4 snapshots in the interaction: 55 documents summarized for January-February, 28 for March-April, 33 for June-July, and 188 for November-December. The document count as well as other sliders and descriptive information are part of the full collage interface, which has been cropped in Figure 3 to only show the timeline plot with images, common phrases and most frequent people named entities across 4 snapshots.

Additional information is displayed in the collage as the user moves the mouse pointer over the representative images. For a given area on the timeline, a number of documents may be represented by a single image, which currently is drawn from the highest scoring shot within the highest scoring document plotted to that area, subject to additional filters (e.g., preference of non-anchorperson shots over anchorperson shots if available). An area of active research will be to evaluate alternate representative image strategies, likely to be informed by user preferences, such as emphasizing images with or without people, close-ups, or overlaid text. As the user mouses over a representative image, tooltips text is displayed indicating the number of documents represented by that image. Additional text can be shown with the tooltip under user control as to the category, count, and thresholds for inclusion; Figure 3's tooltips show up to 4 most frequent locations mentioned in 2 or more documents under the mouse focus.

The user controls how large or small to make the representative images through simple mouse or keyboard control. The January-February timeline shows images at 1/8 resolution in each dimension (the smallest size we have found useful for general visual communication in the collage), while the other timelines in Figure 3 show 1/4 resolution in each dimension, from their original MPEG-1 encoded size. Keeping the image size under user control allows quick inspection of visual details when desired, as well as quick collapsing of images to reduce overlap. In addition, the user can choose to plot a maximum number of images in the collage, in which case after that number has been plotted, remaining documents are represented as squares (or colored geographic areas in the map collage).

Dynamically generating collages through sliders allows for user-directed summarization. In the case of Figure 3, the user is interested in determining the range of stories over time concerning the area around Israel. In January-February, the focus is on Israeli elections, with Barak losing to Sharon. In March-April the focus shifts to armed conflict.

The pictorial overviews add detail. The sunrise shot in June-July can be selected to play its associated story about reduced tourism in Jerusalem due to recent bombings in Israel, and shots of Arafat and smoking buildings in December indicate heightened tension in the area.

The user can select one or more phrases from the text lists and look at only those video documents. For example, selecting "Zinni" and clicking on a "Show Details…" button elsewhere in the collage brings up a list of 12 documents pertaining to U.S. Envoy Zinni during Nov.-Dec. from the initial 409 documents returned for the map query. Selecting "Israeli Prime Minister ariel sharon" produces 7 documents. Selecting both this and "zinni" displays either 1 document or 18, depending on whether the "match ALL" or "match ANY" filtering option is in effect.

All the above dynamic manipulations are carried out instantly in the Informedia digital video library interface, which provides the user an immediate sense of details-on-demand.

There are obvious improvements to be made with these collages. The redundancy within and between lists of named entities and phrases should be removed. Consistent, improved use of upper and lower case should be done across all terms. User studies through heuristic evaluation, think-aloud protocols and other methods (Hearst 1999) should be applied to understand how option selection can be improved, how browsing can be better facilitated, and how the collage can be made into a better facilitation and briefing tool.

Given that the collage is automatically generated from automatically derived metadata containing errors, the collage itself will have shortcomings. Examining such shortcomings can point to areas worthy of further research in the content-based indexing of news. For example, "Beijing TV" is listed as a common phrase for the June-July collage in Figure 3 due to 3 video documents each containing both a story about China via Beijing TV and a separate story about Israel. This mismatch between true story boundaries and automatically derived video document boundaries leads to errors in the collage, and points to the importance of improving automatic video segmentation. As this is a difficult research area, compromises should also be made in the collage generation process to tolerate segmentation errors, e.g., include text terms in summary lists only if they are repeated across a minimum number or minimum percentage of video documents.

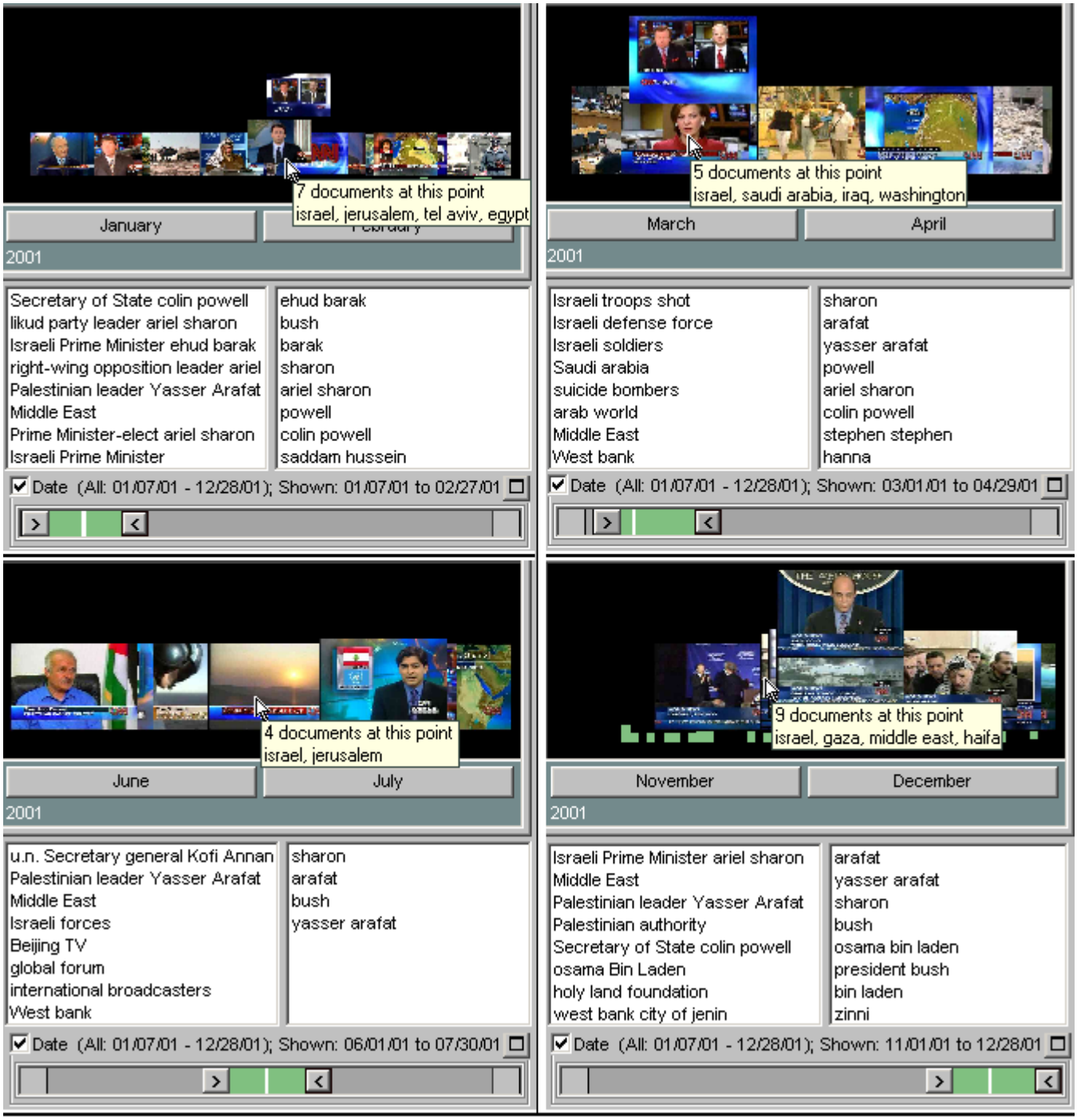

Figure 3. Four-part series showing use of date slider to change the view in the timeline collage, with most common phrases and people listed for all the documents in that date view (tooltips shows locations for the subset plotted at the mouse location).

Preliminary Study on the Utility of Collages As Summaries

To gain insight into the effectiveness of collages as a summary interface, we compared the textual components of our automatically produced collages for our CNN 2001 news collection to summaries of 2001 news material produced elsewhere. The infoplease.com site for "2001 People in the News" lists 83 people who were noteworthy for activity in 2001, providing a brief text description page for each (Infoplease.com 2002). We treated these 83 pages as truth data to measure against the textual information

coverage of our collages and other information sources. We used the 83 names as queries to our corpus, finding 65 that returned one or more video documents (some shown in Figures 4-7). Turning to a biography site, who2.com (Who2.com 2002), we found that only 20 of these 83 names were matched there, indicating the difficulty of coming up with an agreed set of famous people. To complete comparisons, we ran the 83 names as queries to the Google search engine (Google 2002) and used the top-scoring web page outside of who2.com and infoplease.com as another source of data.

In order to make all sources be comparable, we applied the same phrase extraction procedure mentioned in Section 3.2 to three other sources. Our evaluation procedure compares all of the phrases present in the infoplease pages to the who2 and Google-selected pages. We also contrast the infoplease pages with the top ten people, organizations, locations, and common phrases occurring across the most video documents returned from a query on the person's name. To circumvent the issue of scoring partially matched phrases, we compared text at the word-level rather than at the phrase-level by extracting a unique word set from each set of phrases.

Figure 4. Collage for "Jessie Arbogast" from 2001 news set.

The collage output does include images as well as text, as shown in Figure 4 for "Jessie Arbogast." This collage compares well to the Who2 "best known as" summary stating "The boy whose arm was bitten off by a shark." Thus, while only the collage's text is being compared in this investigation, a collage also provides the benefit of imagery to show further context, additional text via the image, and often the person's picture, as in the boy's picture in Figure 4.

The infoplease summary text T_I was taken as truth for each person, with the text for the collage, who2, and Google-located page sources (T_C, T_W and T_G respectively) evaluated against T_I using the F1 metric commonly used in information retrieval (Van Rijsbergen 1979), where F1 = (2 * *precision* * *recall*) / (*precision* + *recall*). As an example, for collage text T_C, *precision* = (words common to T_C and T_I) / (word count in collage text T_C) and *recall* = (words common to T_C and T_I) / (word count in "truth" text T_I). If a source were in perfect agreement with the infoplease summary T_I, its F1 score would be 1.

The F1 scores are reported in Table 1. The relative differences in scores between the collage, who2, and Google-located page sources shows that the collage summary is as good as these other information sources in reporting the text information presented in the infoplease.com noteworthy 2001 people list.

	Data Set	Avg. # Words	Avg. % Recall	Avg. % Prec.	F1
20 names common to 3 sources (avg. word count in T_I is 29.7)	Collage	41.2	30.3	23.6	26.5
	Who2	74.0	41.3	15.8	22.8
	Google-located page	233.0	35.1	6.5	10.9
65 names common to 2 sources (avg. T_I word count is 30.3)	Collage	42.1	29.9	22.6	25.8
	Google-located page	218.8	40.6	8.2	13.7

Table 1. Average recall, precision, and F1 scores compared to infoplease.com "2001 People in the News" summaries

When considering the 20 people matched by all of the sources, the who2 biographical sketch has fair recall, but lower precision because it brings in details before 2001 that are not covered in the infoplease.com summaries. The page returned by Google tends to be much more verbose, sacrificing precision accordingly. The collage text outperforms both in the F1 score. When considering the 65 people matched by the collage and Google-located page,

the collage is more focused and terse with a resulting higher precision and F1 score.

Hence, the evidence suggests that the collage is more valuable as a summarizer for finding information about why a name is in the news than looking at the top-ranked page or going to the biography site. Through the date slider, a collage can be set to a particular time focus such as the year 2001, while web sites cover material across a broad time range at the expense of precision.

Beyond the utility of their text contents as summaries, collages also offer the benefits of folding in imagery and layout, such as seeing a head shot of "Robert Hanssen" in Figure 5. Layout communicates information as well, e.g., seeing the bounds of story reports (February, June) on Nkosi Johnson in Figure 6, and seeing the density of reports on Dennis Tito in May in Figure 7, or seeing geographic distributions via map collages. A straightforward comparison of the collage's text indicates its value as a briefing tool, a value that is enhanced through additional information within the collage. In addition, the collage's interactive nature supports dynamic exploration and revealing of further detail in subsets of interest.

Figure 5. Collage for "Robert Hanssen" from 2001 news set.

Figure 6. Collage generated from 4 video documents returned from "Nkosi Johnson" query (y-axis on timeline is relevance).

Figure 7. Collage generated from 20 video documents returned from "Dennis Tito" query.

Conclusions and Future Work

As the previous sections detailed, collages in their current form are already useful summaries, built dynamically based on automatically generated data and the user's query context. These collages further allow for interactive browsing and exploration. We see great potential for improving collages as both a viewable and interactive summary of video content. The collage as a single, viewable representation of information can be greatly enhanced by incorporating expertise from the field of information design. Font, color, layout, and representation choices all contribute to the effectiveness of collages as used in Section 5, where the collage could be printed out and used effectively as briefing material regarding vast amounts of video. The work of Edward Tufte discusses techniques for improving the persuasive power and coherence of such presentations. In addition, given that the source material is video, collapsed video and audio snippets may enhance the collage's value as briefing material, where the collage can be played as an "auto-documentary" covering a set of video material. Earlier work on video skims (Wactlar et al. 1999) contributes to such an effort.

The dynamics of a collage can be improved by running investigations into their use as direct manipulation interfaces supporting information visualization and exploration. Using qualitative measures such as think-aloud protocols, we can better determine why users make use of some features while ignoring others, providing clues into understanding differences in quantitative metrics. These quantitative metrics include time required to learn the system, time required to achieve goals on benchmark tasks, error rates, and retention of the use of the interface over time (Hearst 1999).

Finally, components feeding into the collages can be improved. Image selection for the collages was based primarily on the correlation of video shots to a user's query. With news video, such a strategy is workable because the audio narrative is closely synchronized to the

visual contents, e.g., the word "volcano" is mentioned as volcano shot is being shown. With other video genres, the audio narrative does not describe the visual contents in a tightly synchronized manner (Li 2000). Hence, much work needs to be done to generate collages outside of news.

Even for news, more information about the images themselves can produce more effective collages. For example, a user interested in key people may want to show only close-up faces, while a different user may want a summary of the same video to emphasize nature setting shots. Formal evaluations into the contribution of imagery to the collage need to be run.

Our discussion of video collages has triggered many other suggestions as to potential enhancements. Perhaps the images should be sized according to their relevance or importance, as is done in the Video Manga interface (Boreczky et al. 2000). Perhaps they should be animated to uncover and then re-obscure themselves when stacked over a common area, as is an option in other visualization systems. These and other interface capabilities will be tested and incorporated as appropriate in our quest to make collages an efficient, effective summary for large video sets.

Acknowledgements. This material is based on work supported by the National Science Foundation (NSF) under Cooperative Agreement No. IRI-9817496. This work is also supported in part by the advanced Research and Development Activity (ARDA) under contract number MDA908-00-C-0037. More details about Informedia research can be found at http://www.informedia.cs.cmu.edu.

References.

Ahlberg, C., and Shneiderman, B. 1994. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays, in Proc. ACM CHI '94 (Boston MA, April 1994), 313-317. ACM Press.

Bates, M.J. 1989. The design of browsing and berrypicking techniques for the on-line search interface. *Online Review* 13(5):407-431.

Bikel, D. M.; Miller, S.; Schwartz, R.; and Weischedel, R. 1997. Nymble: a high-performance learning name-finder, in Proc. 5th Conf. on Applied Natural Language Processing *(ANLP)* (Washington DC, April 1997), 194-201.

Boreczky, J., Girgensohn, A., Golovchinsky, G., and Uchihashi, S. 2000. An Interactive Comic Book Presentation for Exploring Video, in Proc. ACM CHI '00 (The Hague, Netherlands, April 2000), 185-192. ACM Press.

Brill, E. 1992. A simple rule-based part of speech tagger, in Proceedings of the Third Conference on Applied Natural Language Processing (Trento, Italy, 1992), 152-155. ACL.

Chang, S.-F., moderator. 1999. Multimedia Access and Retrieval: The State of the Art and Future Directions, in Proc. ACM Multimedia '99 (Orlando FL, Nov. 1999), 443-445. ACM Press.

Christel, M. 1999. Visual Digests for News Video Libraries, in Proc. ACM Multimedia '99 (Orlando FL, Nov. 1999), 303-311. ACM Press.

Christel, M. and Warmack, A. 2001. The Effect of Text in Storyboards for Video Navigation, in Proc. IEEE ICASSP (Salt Lake City UT, May 2001), vol. III, 1409-1412.

Clarkson, P. and Rosenfeld, R. 1997. Statistical language modeling using the CMU-Cambridge toolkit, in Proc. Eurospeech '97 (Rhodes, Greece, Sept. 1997), 2707-2710. Int'l Speech Communication Assoc.

Ding, W.; Marchionini, G.; and Soergel, D. 1999. Multimodal Surrogates for Video Browsing, in Proc. ACM Conf. on Digital Lib. (Berkeley, CA, Aug. 1999), 85-93. ACM Press.

Google search engine (as of April, 2002), © 2002 Google, http://www.google.com.

Hearst, M.A. 1999. User Interfaces and Visualization. In *Modern Information Retrieval*, edited by R. Baeza-Yates and B. Ribeiro-Neto, New York: Addison Wesley/ACM Press.

Infoplease.com. 2001 People in the News, © 2002 Learning Network, http://www.infoplease.com/ipa/A0878485.html.

Large, A.; Beheshti, J.; Breuleux, A.; and Renaud, A. 1995. Multimedia and Comprehension: The Relationship among Text, Animation, and Captions. *J. American Society for Information Science* 46(5) (June 1995):340-347.

Li, F.; Gupta, A.; Sanocki, E.; He, L.; and Rui, Y. 2000. Browsing Digital Video, in Proc. ACM CHI '00 (The Hague, Netherlands, April 2000), 169-176. ACM Press.

Merlino, A.; Morey, D.; and Maybury, M. 1997. Broadcast News Navigation using Story Segmentation, in Proc. ACM Multimedia '97 (Seattle WA, Nov. 1997), 381-391. ACM Press.

Miller, D.; Schwartz, R.; Weischedel, R.; and Stone, R. 1999. Named Entity Extraction for Broadcast News, in Proc. DARPA Broadcast News Workshop (Washington DC, March 1999), http://www.nist.gov/speech/publications/darpa99/html/ie20/ie20.htm.

Shneiderman, B. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. HCI Lab, Inst. Systems Research, Inst. Advanced Computer Studies, Dept. of Computer Science Tech. Report CS-TR-3665, Univ. of Maryland.

Van Rijsbergen, C. J. 1979. *Information Retrieval.* London: Butterworths.

Wactlar, H.; Christel, M.; Gong, Y.; and Hauptmann, A. 1999. Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library. *IEEE Computer* 32(2) (Feb. 1999): 66-73.

Who2.com. Find Famous People Fast!, © 2002 by Who2?, http://www.who2.com.

Zhang, H.J.; Gong, Y.H.; Smoliar, S.W.; and Yan, S.Y. 1994. Automatic Parsing of news video, in Proc. IEEE Conf. on Multimedia Computing and Systems (Boston MA, May 1994), 45-54.